D196–D201 *Nucleic Acids Research, 2008, Vol. 36, Database issue*
doi:10.1093/nar/gkm980

MIPS: analysis and annotation of genome information in 2007

H. W. Mewes[1,2,*], S. Dietmann[1], D. Frishman[2], R. Gregory[1], G. Mannhaupt[1], K. F. X. Mayer[1], M. Münsterkötter[1], A. Ruepp[1], M. Spannagl[1], V. Stümpflen[1] and T. Rattei[2]

[1]Institute for Bioinformatics (MIPS), German Research Center for Environmental Health, Ingolstaedter Landstraße 1, D-85764 Neuherberg and [2]Technische Universität München, Chair of Genome Oriented Bioinformatics, Center of Life and Food Science, D-85350 Freising-Weihenstephan, Germany

Received September 15, 2007; Revised October 17, 2007; Accepted October 18, 2007

ABSTRACT

The Munich Information Center for Protein Sequences (MIPS-GSF, Neuherberg, Germany) combines automatic processing of large amounts of sequences with manual annotation of selected model genomes. Due to the massive growth of the available data, the depth of annotation varies widely between independent databases. Also, the criteria for the transfer of information from known to orthologous sequences are diverse. To cope with the task of global in-depth genome annotation has become unfeasible. Therefore, our efforts are dedicated to three levels of annotation: (i) the curation of selected genomes, in particular from fungal and plant taxa (e.g. CYGD, MNCDB, MatDB), (ii) the comprehensive, consistent, automatic annotation employing exhaustive methods for the computation of sequence similarities and sequence-related attributes as well as the classification of individual sequences (SIMAP, PEDANT and FunCat) and (iii) the compilation of manually curated databases for protein interactions based on scrutinized information from the literature to serve as an accepted set of reliable annotated interaction data (MPACT, MPPI, CORUM). All databases and tools described as well as the detailed descriptions of our projects can be accessed through the MIPS web server (http://mips.gsf.de).

INTRODUCTION

Both highly curated, specialized data sets as well as automatically generated comprehensive data collections represent the essential backbone of today's genome informatics. At the same time, we need to move towards the implementation of new types of data and enable high-level semantic integration of biological information. MIPS is pursuing these directions by (i) concentrating on expert curation of selected genomes and interaction maps of mammals, fungi and plants, (ii) generating and regularly updating comprehensive, large-scale annotation of protein sequences and their functional classification (SIMAP, PEDANT) and (iii) developing new resources for the semantic representation of heterogeneous and distributed data (CABiNet, NGFN-Portal, GeneSets).

An overview of the general organization and annotation of genome-related information at MIPS is shown in Figure 1 and Table 1.

FUNGAL MODEL ORGANISMS AND KNOWLEDGE TRANSFER TO PATHOGENIC SPECIES

Highly reliable manual annotation of several complete fungal model genomes is a key resource for the fungal research community. Information from well-investigated fungi such as yeast serves often as reference for research in higher eukaryotes. The comprehensive yeast genome database (CYGD) focuses on the most recent chromosomal genome sequence and all annotated genomic features of the baker's yeast *Saccharomyces cerevisiae* (1). The functional annotation was performed using the Functional Catalogue (FunCat), a well-established hierarchical classification system; it is regularly updated (2). In addition, all results of bioinformatics methods applied to the protein sequences of the fungal genomes are part of the GenRE database and can be queried using logical operators. GBrowse (3) is integrated as a graphical interface for displaying the chromosomes and annotation features such as ESTs. For the second model fungus, the orange bread mold *Neurospora crassa*, genome annotation and analysis was also conducted by MIPS (4,5). The presentation of the data in GenRE enables comparative analysis among fungal genomes.

Recently, two well-studied fungal model systems of plant pathogenic species have been annotated. Manual annotation of the severe plant pathogen *Fusarium graminearum* focuses on factors involved in the

*To whom correspondence should be addressed. Tel: +49 89 3187 3580; Fax: +49 89 3187 3585; Email: w.mewes@gsf.de

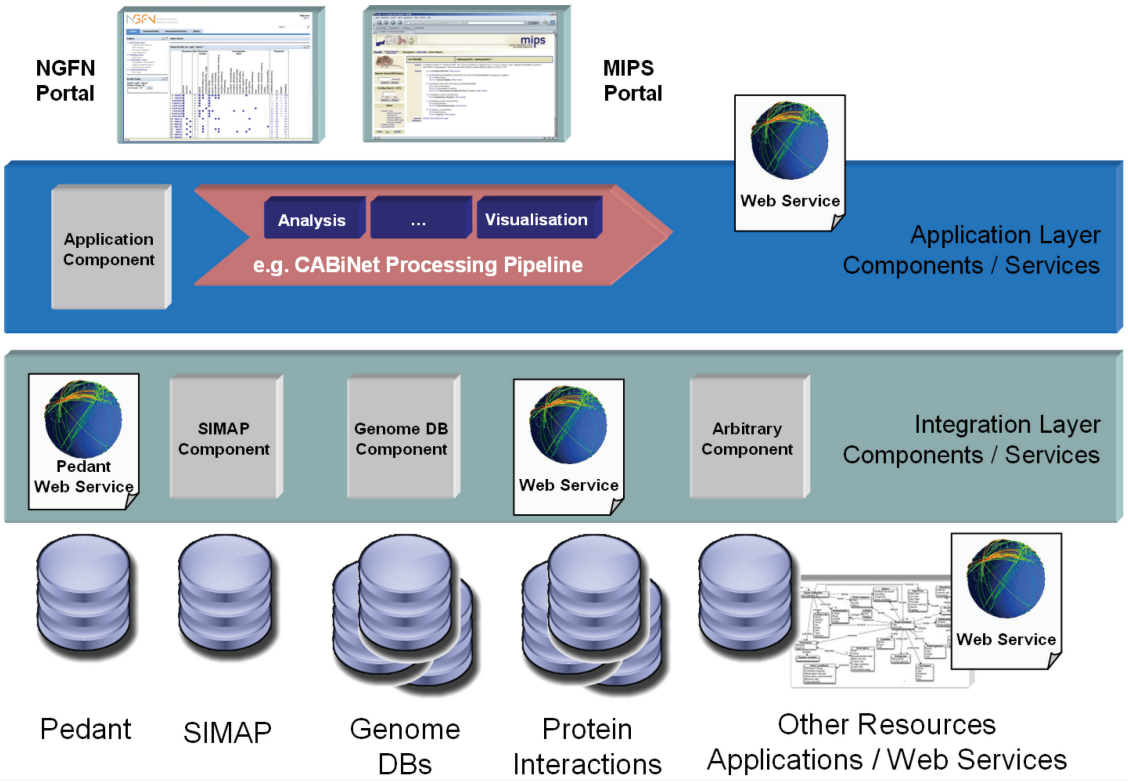

Figure 1. MIPS data sources are heterogeneous by nature. An integration layer based on reusable distributed components and Web Services integrates internal and external sources. These components can be used in the same way for any arbitrary combination of databases and services. Workflows can be generated; for instance a Yeast PPI network can be selected from the interaction database, subsequently the PEDANT system is invoked to retrieve functional annotation and using the SIMAP similarity system the information is mapped onto another genome to predict a putative functional network.

Table 1. URL addresses for MIPS database resources ('http://' is not shown in the table)

Project description	Link
Project overview	mips.gsf.de/projects/
Fungal genome analysis	mips.gsf.de/projects/fungi/
Genome Databases	
CYGD: Baker's Yeast *Saccharomyces cerevisiae*	mips.gsf.de/genre/proj/yeast/
MNCDB: *Neurospora crassa*	mips.gsf.de/genre/proj/ncrassa/
FGDB: *Fusarium graminearum*	mips.gsf.de/genre/proj/fusarium/
MUMDB: *Ustilago maydis*	mips.gsf.de/genre/proj/ustilago/
MATDB: *Arabidopsis thaliana*	mips.gsf.de/projects/plants/thal
MosDB: Rice Genome Database	mips.gsf.de/projects/plants/rice/
URMELDB: European Medicago and Legume Database (*Medicago*)	mips.gsf.de/projects/plants/medicago/
The Lotus Genome Database (*Lotus japonica*)	mips.gsf.de/proj/plants/lotus/
The Environmental Chlamydia Genome Database (*Protochlamydia amoebophila*)	mips.gsf.de/genre/proj/uwe25/
Protein interactions and protein complexes	
MPact: Representation of yeast interaction data	mips.gsf.de/genre/proj/mpact/
MPPI: Mammalian Protein–Protein Interactions	mips.gsf.de/proj/mppi/
MPCDB: Mammalian Protein Complex Data	mips.gsf.de/genre/proj/mpcdb/
Sequence analysis, functional classification, functional modules, genome and network analysis, web services	
SIMAP: Similarity Matrix of Proteins	mips.gsf.de/genre/proj/simap/
Complete Genomes (PEDANT server)	pedant.gsf.de/
FunCat: Functional Catalogue of Proteins	mips.gsf.de/projects/funcat/
CABiNet: Comprehensive Network Analysis	mips.gsf.de/genre/proj/CABiNet/
GenRE: Genome Research Environment	mips.gsf.de/genre/proj/
Genesets: Human functional modules	mips.gsf.de/genre/proj/genesets/
NGFN-Portal: Integration of NGFN data	ngfn.portal.de
MIPS services	mips.gsf.de/webservices/

host–pathogen interaction (6). So far the gene structure of one-third of the ~14 000 protein coding have been manually examined and represented in the GenRE genome database (7). In addition, information on available mutant strains is integrated directing researchers to existing resources and thus avoiding redundant laborious experiments. Localization information of single probes included in the Fusarium Affymetrix GeneChip is listed and also displayed in the graphical interface for detailed validation (8).

In a second fungal pathogen genome annotation project, extensive annotation across the complete genome of the biotrophic smut fungus *Ustilago maydis* resulted in the first comprehensive resource on a basidiomycete fungus. Using this resource, clusters of putative secreted proteins have been revealed which are involved in the virulence of this organism (9). In an ongoing process of comparative annotation of closely related species, the gene models of *U. maydis* as well as the DNA sequence are further refined and build the basis for an upcoming comparative genome resource.

Beyond these four fungal projects, analyses on currently 40 fungal public genomes are maintained in the PEDANT databases, allowing for queries on functional and structural properties such as protein domains. This resource will be continuously extended as new fungal genomes become available and integrated using TopicMaps as a semantic integration platform (www.topicmaps.org).

MIPSPLANTSDB: PLANT DATABASE RESOURCE FOR INTEGRATIVE AND COMPARATIVE PLANT GENOME RESEARCH

Comprehensive and structured information resources are needed to communicate genome data and associated information for model plants and crops. The increase in the available plant genomic data enables powerful comparative analysis and integrative approaches. PlantsDB aims to provide data and information resources for individual plant species to build a platform for integrative and comparative plant genome research. PlantsDB is constituted by the genome databases for *Arabidopsis*, *Medicago*, *Lotus*, rice, maize and tomato. Regular updates and inclusion of latest sequence updates are carried out. Complementary data compilations for *cis*-elements, repetitive elements and extensive cross-species comparisons are implemented.

Beside completely sequenced plant genomes, numerous plant genome-sequencing projects are rapidly progressing. In the near future a bouquet of plant genomes, both models and crops, will be fully available. The availability of a large evolutionary diverse set of plant genomes creates unprecedented opportunities for detailed comparative analysis among different species. PlantsDB aims to address this task by applying a generic though highly flexible modular database infrastructure for a wide range of plant genomes.

The scope of PlantsDB extends beyond individual organism databases. We also aim at providing resources that span multiple species and support specific tasks in comparative and integrative plant genomics, including repeat catalogs, functional classification systems for all plant species, comparative views and search options and a *cis*-element database based on comparative sequence analysis. PlantsDB can be accessed at: http://mips.gsf.de/projects/plants.

PlantsDB organism components

Under the umbrella of PlantsDB, several individual species databases are available. While the individual databases are physically separate, database structures are identical and user interfaces and services provided are similar. This is a prerequisite for easy and intuitive navigation as well as for comparative studies. Due to this generic and modular architecture, new and upcoming plant genome databases can be rapidly implemented.

PlantsDB: data access and retrieval

The MIPS plant genome resources provide access to all genomes covered using common formats and similar interfaces. For every species, all genetic elements are associated with a detailed report of analytical results. Cross-references enable access to associated entries in external databases. Visualization of genetic elements and genomic context is undertaken via a GBrowse Genome Browser interface (3); queries across individual or multiple databases can be performed using different search forms and search engines. In addition, ftp access to various data forms is provided. Beside sequence files for all contigs and protein-coding genes, this section offers the functionality to create and download user-defined Genome Annotation Markup Element files (GAMEXML) used by the Apollo Genome Browser. Apollo provides a detailed graphical viewer for genome data with more flexible interaction possibilities than a browser-based display (10).

The MIPS Repeat Element Database (mips-REdat) and the MIPS Repeat Element Catalog (mips-REcat)

Plants genomes are notorious for containing large amounts of repetitive elements. Their repeat content ranges from 20% to 30% in *Arabidopsis* to over 90% in large genomes such as wheat (16 000 Mb). Evidence is accumulating that mobile elements contribute to genetic diversity by shuffling functional blocks and in genome analysis repeat elements can be used as fossil traces to study genome evolution (11–13). An important prerequisite for such comparative analyses is the consistent cross-species annotation and classification of repeat elements. For this purpose, we developed a database of plant repeats elements (mips-REdat) and a generic repeat classification schema (mips-REcat). While Mips-REdat represents an exhaustive collection of plant repetitive elements, Mips-REcat has been designed for automated repeat element annotation and flexible data retrieval. It integrates and

extends existing repeat classifications into a systematic hierarchical tree structure.

MotifDB *cis*-element database

One of our research goals is the comprehensive discovery of transcription factor binding sites (TFBs) in plant genomes to enable and stimulate the study of regulatory networks in plants. Based on a combination of phylogenetic footprinting and motif discovery in promoters of co-expressed genes, we determine candidate *cis*-regulatory elements (14,15). Results of this study have been integrated in the comparative section of PlantsDB (http://mips.gsf.de/proj/plant/webapp/expressionDB/index.jsp). Queries for individual genes of interest, for co-expressed genes, for candidate TFBs and for specific biological processes and categories of interest are supported.

PEDANT AND SIMAP AS COMPREHENSIVE RESOURCES FOR PROTEIN ANNOTATION

MIPS has established a computational pipeline based on two major software tools—SIMAP (16) and PEDANT (17)—which serves as the central backbone for processing sequence information. The main intention is to automate and streamline the annotation process, reduce its computational cost and avoid duplication of effort. The SIMAP database provides the pre-calculated sequence similarity network and annotation features such as InterPro domains for all proteins contained in the major public sequence databases. As of September 2007, SIMAP covers about 17 million proteins and more than 6 million non-redundant sequences. It also provides the total of 560 Mio. InterPro domain assignments in 5.7 Mio. proteins. Due to the incremental update capabilities of SIMAP, repeated or redundant calculations are avoided.

PEDANT is a genome analysis software suite designed for unsupervised annotation of genomes. The new version of PEDANT (PEDANT3) represents a comprehensive Java workbench, which covers diverse aspects of genome annotation—from gene prediction to functional and structural characterization of proteins and genome comparison. PEDANT3 is directly integrated with SIMAP from which it can import any up-to-date pre-calculated results. Only those features that are not covered by SIMAP get calculated by PEDANT's own workflow-based processing engine. The current version of the PEDANT genome database makes the annotation of nearly 500 genomic sequences available.

REPRESENTATION OF HETEROGENEOUS AND DISTRIBUTED DATA FOR NGFN

The compilation of various distributed data is only one step towards biological knowledge bases allowing for seamless access to information across many independent data sources. The integration of heterogeneous biological information is indeed a daunting task in view of the sheer amount of data. A promising perspective is opened by the use of semantic information systems allowing for associative linking of independent sources. While some minor examples for the installation of semantic technologies based on the World Wide Web Consortium's (W3C) RDF technology (http://www.w3org/RDF/) exist, the breakthrough of this advanced technology is still hold back since suitable community agreements covering the semantic description of even the most important biological information types are missing.

For the NGFN (Germany's National Genome Research Network) independent heterogeneous experimental data and associated information had to be integrated. Within the hundreds of projects of the NGFN, a wide variety of heterogeneous information exists which is typically connected to (sets of) genes. We implemented an intermediate approach which was inspired by two of the successful Web 2.0 approaches: (i) the tagging of information (uniquely describing what a single piece of information is about) and (ii) the aggregation of heterogeneous content from different resources within a web page. Content aggregation is implemented with a so-called Portal solution connecting content from different resources by portlets. Portlets are independent components on a web page capable to access and render independent content. For example, this technology enables us to retrieve information related to some gene(s) within one portlet from a local database while a second one accesses an external Web Service. Placing both on one common web page provides a seamless view on different resources. The second concept (tagging) is semantically rather weak because it does not describe the context of an individual tag. Thus, this part was extended with a typed tagging mechanism allowing the user to describe the meaning of a specific tag. The NGFN portal allows for typed tagging of different levels of information. It starts with tagging general aspects of an investigation, followed by tagging of individual related results and finally ending in specific tags for single genes being part of a result. For convenience, we added certain predefined tags e.g. reflecting the functional annotation of genes as well as their occurrence in literature or disease databases.

GENESETS

The GeneSets project aims to provide a platform for compiling and annotating functional modules in human and mammalian species. The project involves (i) dynamically collecting all publicly available protein–protein interaction data sets, (ii) providing a score for each interaction using a fully connected Bayesian network approach and considering manually curated protein complexes (CORUM, this issue), (iii) integrating evolutionary and functional protein profiles in a flexible manner and (iv) extracting modules from the interaction network by an algorithm that exhaustively enumerates all densely connected modules (Georgii *et al.*, manuscript in preparation). Furthermore, the integration of orthology

profiles and tissue expression profiles as constraints during the module enumeration process refines the internal sub-modular structure and yields, in many cases, a straightforward functional interpretation on essential or peripheral proteins providing a useful context for the analysis of mouse gene knock-out or human disease mutation data. In addition, to uncover novel functional modules, the GeneSets database serves as a resource for the local analysis of known protein complexes (CORUM, this issue) and more general functional modules, as defined by human experts, such as those provided by the FunCat (2) and KEGG databases (18). Visualization of modules is performed by the CABInet suite (19). The collection of modules is externally accessible for enrichment analysis of experimental gene expression or proteomics data sets. The GeneSets resource thus complements related analysis suites that are primarily based on manual functional classifications, such as the Gene Ontology or the KEGG database, by pre-compiling and condensing the vast amount of publicly available experimental information on functional relationships between genes.

SUMMARY

The classical concept of collecting sequence-related information linked to individual genetic elements such as proteins needs to be revised to cope with the requirements of systems biology. Today, annotating genomes increasingly means characterizing functional modules and networks for modeling cellular processes. Overall, while the collection of nucleic acid and protein sequence data has become an industrial workflow, the transition from 'parts lists' to functional interaction graphs, the 'wiring diagrams' is still subject of research and has not yet reached the level of a unifying concept (20). To understand complex problems such as the role of genetic and environmental factors in human disease, additional dimensions of the data need to be considered to represent cellular state models in space and time. These complex relations need to be reflected in suitable standardized data structures and visualization tools for the representation of multi-dimensional interaction data.

The value of public genome databases strongly depends on the quality of the information associated to the sequence and the ease of access to structured consistent and reliable data. The number of genomes being investigated in detail will remain low in comparison to the total number of sequenced genomes including metagenomes, which will increase with dramatic speed. MIPS follows a two-way strategy to contribute to the wealth of genome databases. On the one hand, reference genomes and interaction databases are maintained by careful manual inspection. On the other hand, comprehensive methods to transfer information to all genomes available in a systematic way are applied at large scale.

ACKNOWLEDGEMENTS

This work was supported by the Federal Ministry of Education, Science, Research and Technology (BMBF: BFAM: 031U112C, GABI: 0312270/4, NGFN: 01KW9710), the Deutsche Forschungsgemeinschaft (Bioinformatics Munich, BIM), and the Max-Planck-Society (MUMDB). We thank Biomax Informatics AG for providing the PEDANT Genome Annotation System. Funding to pay the Open Access publication charges for this article was provided by the GSF—National Research Center for Environment and Health.

Conflict of interest statement. None declared.

REFERENCES

1. Güldener,U., Münsterkötter,M., Kastenmüller,G., Strack,N., Van Helden,J., Lemer,C., Richelles,J., Wodak,S.J., Garcia-Martinez,J. *et al.* (2005) CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res.*, **33**(Database Issue), D364–D368.
2. Ruepp,A., Zollner,A., Maier,D., Albermann,K., Hani,J., Mokrejs,M., Tetko,I., Guldener,U., Mannhaupt,G. *et al.* (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.*, **32**, 5539–5545.
3. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
4. Schulte,U., Becker,I., Mewes,H.W. and Mannhaupt,G. (2002) Large-scale analysis of sequences from *Neurospora crassa*. *J. Biotechnol.*, **94**, 3–13.
5. Galagan,J.E., Calvo,S.E., Borkovich,K.A., Selker,E.U., Read,N.D., Jaffe,D., FitzHugh,W., Ma,L.J., Smirnov,S. *et al.* (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature*, **422**, 859–868.
6. Cuomo,C.A., Güldener,U., Xu,J.R., Trail,F., Turgeon,B.G., Di,P.A., Walton,J.D., Ma,L.J., Baker,S.E. *et al.* (2007) The Fusarium graminearum genome reveals a link between localized polymorphism and pathogen specialization. *Science*, **317**, 1400–1402.
7. Güldener,U., Mannhaupt,G., Münsterkötter,M., Haase,D., Oesterheld,M., Stümpflen,V., Mewes,H.W. and Adam,G. (2006) FGDB: a comprehensive fungal genome resource on the plant pathogen Fusarium graminearum. *Nucleic Acids Res.*, **34**, D456–D458.
8. Güldener,U., Seong,K.Y., Boddu,J., Cho,S., Trail,F., Xu,J.R., Adam,G., Mewes,H.W., Muehlbauer,G.J. *et al.* (2006) Development of a Fusarium graminearum Affymetrix GeneChip for profiling fungal gene expression in vitro and in planta. *Fungal Genet. Biol.*, **43**, 316–325.
9. Kamper,J., Kahmann,R., Bolker,M., Ma,L.J., Brefort,T., Saville,B., Banuett,F., Kronstad,J., Gold,S. *et al.* (2006) Insights from the genome of the biotrophic fungal plant pathogen Ustilago maydis. *Nature*, **444**, 97–101.
10. Lewis,S.E., Searle,S.M., Harris,N., Gibson,M., Lyer,V., Richter,J., Wiel,C., Bayraktaroglir,L., Birney,E. *et al.* (2002) Apollo: a sequence annotation editor. *Genome Biol.*, **3**, RESEARCH0082.
11. Kazazian,H.H.Jr (2004) Mobile elements: drivers of genome evolution. *Science*, **303**, 1626–1632.
12. Frost,L.S., Leplae,R., Summers,A.O. and Toussaint,A. (2005) Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.*, **3**, 722–732.
13. SanMiguel,P., Gaut,B.S., Tikhonov,A., Nakajima,Y. and Bennetzen,J.L. (1998) The paleontology of intergene retrotransposons of maize. *Nat. Genet.*, **20**, 43–45.
14. Haberer,G., Mader,M.T., Kosarev,P., Spannagl,M., Yang,L. and Mayer,K.F. (2006) Large-scale cis-element detection by analysis of correlated expression and sequence conservation between Arabidopsis and Brassica oleracea. *Plant Physiol.*, **142**, 1589–1602.
15. Haberer,G., Young,S., Bharti,A.K., Gundlach,H., Raymond,C., Fuks,G., Butler,E., Wing,R.A., Rounsley,S. *et al.* (2005) Structure

and architecture of the maize genome. *Plant Physiol.*, **139**, 1612–1624.
16. Rattei,T., Arnold,R., Tischler,P., Lindner,D., Stumpflen,V. and Mewes,H.W. (2006) SIMAP: the similarity matrix of proteins. *Nucleic Acids Res.*, **34**, D252–D256.
17. Riley,M.L., Schmidt,T., Artamonova,I.I., Wagner,C., Volz,A., Heumann,K., Mewes,H.W. and Frishman,D. (2007) PEDANT genome database: 10 years online. *Nucleic Acids Res.*, **35**, D354–D357.
18. Kanehisa,M., Goto,S., Hattori,M., oki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
19. Oesterheld,M., Mewes,H.W. and Stümpflen,V. (2007) Analysis of integrated biomolecular networks using a generic network analysis suite. *J. Integr. Bioinformatics*, in press.
20. Frishman,D. (2007) Protein annotation at genomic scale: the current status. *Chem. Rev.*, **107**, 3448–3466.